Bayesian Interim Analysis of Lifetime Data

by

Seymour Geisser and George Papandonatos
University of Minnesota and SUNY at Buffalo

Technical Report No. 619
March 1997

Bayesian Interim Analysis of Lifetime Data [1]

GEORGE D. PAPANDONATOS
Department of Statistics
242 Farber Hall, 3435 Main St.
School of Medicine and Biomedical Sciences
Buffalo, N.Y. 14214-300

and

SEYMOUR GEISSER
School of Statistics, University of Minnesota
270 Vincent Hall, 206 Church St. SE
Minneapolis, MN 55455
6 April, 1997

Abstract

In hypothesis testing involving censored lifetime data that are independently distributed according to an Accelerated Failure Time (AFT) model, it is often of interest to predict whether continuation of the experiment will significantly alter the inferences drawn at an interim stage. Approaching the problem from a Bayesian viewpoint, we suggest a possible solution based on Laplace approximations to the posterior distribution of the parameters of interest and on Markov Chain Monte Carlo. We apply our results to Weibull data from a carcinogenesis experiment on mice.

Running title: Bayesian Interim Analysis.

Key words and phrases: Bayesian Predictive Inference; Data monitoring; Interim Analysis.

AMS 1991 subject classifications: Primary 62F15, 62H15; secondary 62L99; 62N05.

[1] Research partially supported by National Institutes of General Medical Sciences Grant GM-25271.

1 Introduction.

In the field of interim monitoring of clinical trials, a group–sequential alternative to the decision–theoretic fully–sequential procedures of Berry (1985) has been forcefully advocated by Geisser (1992, 1993a) and further developed in collaborative work with Johnson (1993). Noting the difficulties clinical researchers have experienced in specifying an appropriate loss function and defining the relevant patient horizon, Geisser & Johnson recommended that a Bayesian predictive stance be adopted instead, close in spirit to the stochastic curtailment ideas of DeMets & Lan (1984), Choi & Pepple (1989) and Spiegelhalter et al. (1986, 1988). The syncretic approach favored by this latter group - a combination of significance testing at the terminal decision point with Bayesian analysis at the interim point - was found wanting, however, and was discarded in favor of a fully Bayesian treatment of the problem. An overview of recent work on this topic is provided by Geisser (1993b).

The setup of most interest to us is one where a laboratory or regulatory agency is evaluating a new therapy and stipulates that a decision on its effectiveness cannot be made unless data becomes available on at least s subjects, at which point it may be decided that it is at least as effective as some standard, it is found ineffective or it is deemed sufficiently promising to justify further sampling. We assume that the problem can be put in a parametric framework - with $g(\boldsymbol{\theta})$ a scalar function of the parameters indicative of the effectiveness of the treatment - and that a prior can be elicited for $g(\boldsymbol{\theta})$ or an uninformative analysis can be agreed upon. The final decision after seeing data $\mathbf{y}_s$ can then be made dependent upon $P_s = P[g(\boldsymbol{\theta}) \in G \,|\, \mathbf{y}_s]$, the probability content of an appropriately chosen credible set G in the support of $g(\boldsymbol{\theta}) \,|\, \mathbf{y}_s$. The treatment would then be declared effective if P_s exceeded p_2, the trial abandoned if P_s fell below p_1 or a decision withheld for $P_s \in [p_1, p_2]$.

The minimum–sample–size requirement s is assumed to have arisen from a fixed design minimising some preposterior measure of loss, when a reasonable loss function can be specified. Alternatively, s could be the maximiser of the Kullback–Leibler (KL) divergence between the prior and posterior density of the parameters of interest, subject to a fixed budget for the experiment. Proposed by Lindley (1956) and given additional theoretical justification by Bernardo (1979), KL divergence is a natural measure of the information provided by the data and has recently been put forward as a design criterion for clinical trials by Parmigiani & Berry (1994). It is often equivalent to maximising the probability content of select credible sets of the terminal posterior, an ad-hoc criterion one can also use directly as an easier-to-interpret design goal.

Since sampling is costly, the experimenter may want to take a training sample of size $n \leq s$ first and use P_n as a rough guide to what P_s will turn out to be, before entering an additional $m \geq s - n$ subjects into the study. But, the differences between the interim and terminal posterior probabilities may well be substantial and it is more appropriate to examine the predictive distribution of P_s given the interim data. Complicated to start with, the evaluation of this quantity becomes harder if uncertainty about the accrual rate of future entrants to the study is to be taken into account and allowance made for censoring in both

the present and future sample. Fast and accurate approximations to this quantity are thus essential to the implementation of the Bayesian approach. In this paper, we limit ourselves to the implementation and evaluation of Bayesian predictive stopping rules for loglifetime data arising from linear regression models with uninformative censoring and arbitrary, but known, error distributions.

We mention in passing that the Bayesian predictive approach can also be of use in clinical investigations whose purpose is not tied to a specific decision, but which aim rather to "contribute usefully to knowledge", a shift of emphasis designed partly to accommodate Armitage's (1989) criticism of early–stopping rules. In a trial "to learn", a specified increase in information may itself be the primary goal and the aim of the interim analysis would then be to ascertain the probability that this goal will be reached eventually, subject to the aforementioned budget and time limitations.

2 Is a detailed predictive analysis worthwhile?

When the accrual rate is deterministic and observation of all n individuals in the initial sample ceases at the interim point, the quantity of interest to us is of the form

$$P[\,P[\,g(\theta) \in G \,|\, \mathbf{y}_{n+m}\,] \geq p \,|\, \mathbf{y}_n] \quad = \quad P[\,P_{n+m} \geq p \,|\, \mathbf{y}_n], \tag{2.1}$$

Its evaluation at p_1, p_2 allows us to calculate at the interim point the probability that either one of the three possible decisions will eventually be reached after termination of the study. If $P[\,P_{n+m} \geq p_2 \,|\, \mathbf{y}_n\,]$ falls below some lower bound chosen to reflect the losses of correct and incorrect decisions, we may well decide to abandon the trial early, rather than commit more resources in an apparently futile effort to demonstrate a treatment effect. By allowing $m + n$ to exceed the minimum sample size s, we can also conduct a sensitivity analysis of the type described in Geisser (1993a).

Although (2.1) seems a reasonable quantity to examine in this context, its evaluation can sometimes be avoided altogether, since we can bound it both above and below. In order to do this, let us first denote by $(\Omega, \mathcal{F}, Q)$ the common probability space on which $\theta, y_1, y_2, \ldots$ are defined and set $\mathcal{F}_k = \sigma(y_1, \ldots, y_k)$, $\mathcal{F}_\infty = \sigma(U_{k=1}^{\infty} \mathcal{F}_k)$.

If $A = \{\omega : g \circ \theta\,(\omega) \in G\}$, we see that

$$P_{n+m} = P[\,g(\theta) \in G \,|\, \mathbf{y}_{n+m}\,] = E[I_A \,|\, \mathcal{F}_{n+m}\,]$$

is but the conditional expectation of a bounded function of ω. Since

$$E[P_{n+m} \,|\, \mathcal{F}_n] = E[\,E[I_A \,|\, \mathcal{F}_{n+m}\,] \,|\, \mathcal{F}_n] = E[I_A \,|\, \mathcal{F}_n] = P_n, \tag{2.2}$$

the sequence is a martingale and converges a.s. Q and in mean square to $E[\,I_A \,|\, \mathcal{F}_\infty\,]$ as m increases, by sec. VII of Doob (1953). When I_A is $\mathcal{F}_\infty$ measurable, I_A itself is the limit a.s Q. Doob's proof does not make transparent the dependence of the convergence set on the prior. This is clearer in an earlier paper of Doob (1949), which pertains to the more restrictive case

of i.i.d sequences of observations with proper priors on the parameters. There it is seen that the set of θ's for which convergence fails to occur has measure 0 under our choice of prior. Since our prior may assign measure zero to arbitrarily large subsets of Euclidean space, this convergence result is not entirely satisfactory and serves as a reminder to Bayesians that they should act conservatively and assign positive, if negligible, prior probability even to values of the parameter space which they regard as unlikely to occur. An extension of Doob's results to improper priors is given in Hartigan (1983).

Since our best guess of P_{n+m} conditional on $\mathbf{y}_n$ is the current value P_n, enlarging the sample by m additional observations does not guarantee that our goal will necessarily be attained and may actually move us further away from it. However, by Jensen's inequality, P_{n+m}^2 is a submartingale so that

$$Var[\, P_{n+m} \,|\, \mathcal{F}_n] = E[\, P_{n+m}^2 \,|\, \mathcal{F}_n] - P_n^2 \tag{2.3}$$

is nondecreasing in m and is actually strictly increasing when the sequence of conditional distributions is non-degenerate. It is exactly this increase in the variability of the predictive distribution that leads us to be hopeful that the goal may be reached by taking further observations. On the other hand, a bound to the probability of ever attaining our goal is provided by the Markov inequality, which implies that

$$1 - \frac{1 - P_n}{1 - p} \leq P[\, P_{n+m} \geq p \,|\, \mathcal{F}_n] \leq \frac{P_n}{p}. \tag{2.4}$$

For P_n smaller than p, the upper bound becomes sharp. This indicates that we may have little hope of reaching our goal if the interim posterior probability that $g(\theta) \in G$ is way below the threshold value p_2; a quite intuitive result that is also seen to be supported by the theory. Since P_{n+m} tends to I_A a.s. Q as m increases, the limiting predictive distribution of $P_{n+m} \,|\, \mathbf{y}_n$ is supported entirely on $\{0, 1\}$ with masses depending on our current beliefs about $g(\theta)$ being in G, i.e.

$$P[\, P_{n+m} \geq p_2 \,|\, \mathcal{F}_n] \longrightarrow P[I_A \geq p \,|\, \mathcal{F}_n] = P_n, \tag{2.5}$$

as $m \to \infty$, for all $p \in (0, 1)$. The above analysis suggests that there is a lot that we can say about the predictive distribution of $P_{n+m} \,|\, \mathbf{y}_n$, without actually having to evaluate it. The results we will be presenting in the remainder of this paper are then only likely to be useful when P_n is smaller than p but not by much and we are limited in the number of future observations that we can take.

3 Approximating the predictive probability of success

Let us pretend for a moment that the trial has been brought to its conclusion and $\mathbf{y}_{n+m}$ is actually available. The calculation of $P[\, g(\theta) \in G \,|\, \mathbf{y}_{n+m}]$ then simply requires the terminal posterior of $g(\theta)$. However, for censored data problems with nonconjugate priors, obtaining

this posterior in closed form is usually impossible. Although a naive approximation to the posterior could be based on a normal distribution with mean and covariance matrix determined by the first two derivatives of the log–posterior evaluated at its mode, for the sample sizes encountered in practice the normal approximation is likely to be grossly inadequate. Heavy censoring can drastically reduce the effective sample size and result in posterior distributions that exhibit substantial asymmetry. The Laplace approximation, proposed by Tierney & Kadane (1986), improves upon the normal approximation in that it can handle both skewness and kurtosis. Unfortunately, it is sensitive to all but mild multimodality. It can be shown, though, that for the generalised gamma distribution, a family that includes both the lognormal and the Weibull as special cases, there exists a parameterisation in which the log-likelihood is globally strictly concave. So, with flat priors, we need not be overly concerned about multimodality of the posterior, at least for this flexible family of lifetime distributions.

The Laplace posterior requires numerical integration to find its normalising constant, with a separate constrained maximisation at each integration point. Even if the maximisations converge, they tend to be quite costly to compute. What is worse from our point of view, in the usual case where the parameter space Θ is unbounded we have to truncate it for integration purposes to an interval that contains essentially all the probability mass. But, the truncation limits will differ from sample to sample and can only be obtained graphically, by visual examination of the unnormalised posterior. This requires interaction with the user, which one would rather avoid. A reparameterisation may transform Θ to a bounded interval, but the Laplace approximation usually performs best when it is left unrestricted. There is, however, a way out of this conundrum: by careful examination of the error rates, derived in Kass et al. (1990), it is possible to show that the Laplace approximation can be replaced by its Edgeworth–type expansion around the posterior mode with no additional loss of accuracy, at least in $O\left((n+m)^{-1/2}\right)$ neighborhoods of the mode. Once an Edgeworth-type expansion is available, it can be integrated asymptotically and third order approximations to credible intervals, quantiles, moments etc. can be easily found using standard procedures such as Cornish-Fisher inversion. Alternatively, densities from well-known distributions can be fit by matching either the first four derivatives at the mode, or the first four cumulants. Credible intervals would then be immediately available from standard computer packages. The first approach has been advocated by DiCiccio et al. (1990, 1991), while the second is closer to the recommendations of Viveros and Sprott (1987). Papandonatos & Geisser (1997) have examined both approaches in the context of linear regression models with possible censoring and have shown that impressive accuracy can be attained, even for terminal sample sizes as small as fifteen.

At the interim stage only the first n observations will be available to us. The remaining m will have to be simulated from the predictive distribution of $\mathbf{y}_m \mid \mathbf{y}_n$. Since this can be represented as mixture over Θ of the sampling densities of $\mathbf{y}_m \mid \boldsymbol{\theta}$, with the interim posterior $\boldsymbol{\theta} \mid \mathbf{y}_n$ as the mixing measure, one needs to be able to draw samples from $p(\boldsymbol{\theta} \mid \mathbf{y}_n)$. Typically, the subsequent generation of log–lifetimes from $p(\mathbf{y}_m \mid \boldsymbol{\theta})$ is trivial or at least

well-documented in the literature. However, censoring and the possible lack of conjugacy of the prior usually render the posterior intractable and require the use of Markov Chain Monte Carlo techniques, a comprehensive survey of which appears in Tierney (1994). [2]

Once the terminal posterior mode $\hat{\boldsymbol{\theta}}_{n+m}$ has been located, $P[g(\boldsymbol{\theta}) \in G \mid \mathbf{y}_{n+m}]$ can be approximated by expanding the terminal posterior of $g(\boldsymbol{\theta})$ around $g(\hat{\boldsymbol{\theta}}_{n+m})$, as in Papandonatos & Geisser (1997). When the size of the future sample is small relatively to that of the initial sample and stage II is short in duration, we would expect $p(\boldsymbol{\theta} \mid \mathbf{y}_n)$ and $p(\boldsymbol{\theta} \mid \mathbf{y}_{n+m})$ to be similar. In that case, $\hat{\boldsymbol{\theta}}_{n+m}$ will not differ substantially from $\hat{\boldsymbol{\theta}}_n$, the posterior mode at the interim stage, which can then serve as the starting value for the Newton iteration. On the other hand, if m/n is large and stage II is long enough for only a few of the participants to remain alive at the end of the experiment, we would expect $\hat{\boldsymbol{\theta}}_{n+m}$ to be closer to $\tilde{\boldsymbol{\theta}}_m$, the simulated value of $\boldsymbol{\theta}$ that was used to generate both the m lifetimes of the future sample and the latent lifetimes of the stage I survivors. A weighted average of $\hat{\boldsymbol{\theta}}_n$ and $\tilde{\boldsymbol{\theta}}_m$ can then be used to initialise Newton's algorithm, with the weights chosen empirically to reflect both the relative sample sizes and the degree of censoring. We have found that a matrix weighted average with weights equal to the information matrices for the original and future samples, evaluated at the interim posterior mode and the simulated value of $\boldsymbol{\theta}$ respectively, is quite adequate in this respect.

4 Simulating the final stage of trial

For convenience we shall treat the data as representing individual lifetimes and let $\mathbf{T}_n = (\mathbf{T}_d, \mathbf{T}_c, \mathbf{T}_l)$ denote the n-vector of times under observation of the d individuals in the initial sample that died during stage I, those c whose observation ceased at or prior to the interim point and the l survivors that enter stage II respectively. Since we allow staggered entry into the trial, the censoring times may all differ, even if observation of the initial sample ceases at the same time point for all individuals in it. At the end of the trial, the data on which we will base our inferences for $g(\boldsymbol{\theta})$ can be similarly partitioned into $\tilde{\mathbf{T}}_{n+m} = (\mathbf{T}_d, \mathbf{T}_c, \tilde{\mathbf{T}}_l, \tilde{\mathbf{T}}_m)$, where $\tilde{\mathbf{T}}_l$ denotes time under observation during the entire trial for those stage I survivors that remain under observation in stage II and $\tilde{\mathbf{T}}_m$ is the time under observation of the m individuals of the future sample.

It is assumed that the actual lifetimes are conditionally independent given $\boldsymbol{\theta}$ and that the distribution of their natural logarithm can be adequately described by the linear regression model

$$\log L = \theta_0 + \boldsymbol{\theta}_1^T \mathbf{w} + Z \exp \theta_2,$$

with the intercept θ_0, the slope vector $\boldsymbol{\theta}_1$ and the logscale parameter θ_2 making up the

[2] All of the above are predicated on a fixed design on s individuals. In trials "to learn", we would first have to approximate the preposterior distribution of the Lindley Information Measure in order to find s. The same methods used in sampling from $\boldsymbol{\theta} \mid \mathbf{y}_s$ would also be applicable with $\mathbf{y}_s$ drawn from its prior predictive distribution, assuming it is proper.

parameter vector θ. In addition, the error distribution $F_Z(z)$ will be taken as known. An estimate of $P\left[P\left[g(\theta) \in G \,|\, \tilde{\mathbf{T}}_{n+m}\right] \geq p \,|\, \mathbf{T}_n\right]$ can then be obtained as follows:

1. Generate K samples $\tilde{\mathbf{T}}_{n+m}^{(1)}, \ldots, \tilde{\mathbf{T}}_{n+m}^{(K)}$ from the predictive distribution of $\tilde{\mathbf{T}}_{n+m} \,|\, \mathbf{T}_n$.
2. For the k'th sample calculate $P^k = P\left[g(\theta) \in G \,|\, \tilde{\mathbf{T}}_{n+m}^{(k)}\right]$.
3. Return $\sum_{k=1}^{K} I\{P^k \geq p\}/K$, together with an estimate of its standard error.

We will first describe the simulation step in detail. In order to do so, we will employ bracket notation to denote densities, with the joint, conditional and marginal densities of two generic random variables U, V given by $\{U, V\}, \{U \,|\, V\}, \{U\}$ respectively. In addition, the integral sign will imply marginalisation, so that $\int \{U \,|\, V\}\{V\} = \{U\}$. It can then be shown that that the predictive density $\{\tilde{\mathbf{T}}_{n+m} \,|\, \mathbf{T}_n\}$ decomposes into

$$\{\mathbf{T}_d, \mathbf{T}_c, \tilde{\mathbf{T}}_l, \tilde{\mathbf{T}}_m \,|\, \mathbf{T}_d, \mathbf{T}_c, \mathbf{T}_l\} \;=\; \{\mathbf{T}_d \,|\, \mathbf{T}_d\}\,\{\mathbf{T}_c \,|\, \mathbf{T}_c\}\,\{\tilde{\mathbf{T}}_l, \tilde{\mathbf{T}}_m \,|\, \mathbf{T}_d, \mathbf{T}_c, \mathbf{T}_l\}.$$

Here $\{\mathbf{T}_d \,|\, \mathbf{T}_d\}, \{\mathbf{T}_c \,|\, \mathbf{T}_c\}$ are point-mass distributions, while $\{\tilde{\mathbf{T}}_l, \tilde{\mathbf{T}}_m \,|\, \mathbf{T}_d, \mathbf{T}_c, \mathbf{T}_l\}$ equals

$$\int \{\tilde{\mathbf{T}}_m \,|\, \mathbf{L}_m, \tilde{\mathbf{C}}_m\}\{\mathbf{L}_m \,|\, \theta\}\{\tilde{\mathbf{C}}_m\}\{\tilde{\mathbf{T}}_l \,|\, \mathbf{L}_l, \tilde{\mathbf{C}}_l\}\{\mathbf{L}_l \,|\, \theta, \mathbf{T}_l\}\{\tilde{\mathbf{C}}_l \,|\, \mathbf{T}_l\}\{\theta \,|\, \mathbf{T}_d, \mathbf{T}_c, \mathbf{T}_l\},$$

where L is the true lifetime of an individual and $\tilde{C}$ its censoring time at the terminal point. Since we only deal with noninformative censoring, we do not allow the distribution of $\tilde{C}$ to depend on θ. By definition, $\tilde{T} = min(L, \tilde{C})$ is a constant when both $L, \tilde{C}$ are known, so we concentrate on the nondegenerate densities:

1. $\{\mathbf{L}_m \,|\, \theta\}$: If we have access to a statistical package that allows us to sample from the error distribution $F_Z(z)$ directly, we can generate the loglifetimes of the future sample by a location-scale transformation of Z: $\log L = \theta_0 + \theta_1^T \mathbf{w} + Z \exp\theta_2$.
2. $\{\mathbf{L}_l \,|\, \theta, \mathbf{T}_l\}$: Since all stage I survivors were censored at the interim point, we must left-truncate the usual sampling distribution of $\mathbf{L}_l$ at $\mathbf{T}_l$. Although we could generate the individual lifetimes as in (1) and retain only those that satisfy the constraint, it is more efficient to use 1-1 inverse C.D.F. sampling by noting that the individual lifetimes are conditionally independent given θ and satisfy

$$P\,[L \leq t_2 \,|\, L \geq t_1] \;=\; \frac{F_Z(z_2) - F_Z(z_1)}{1 - F_Z(z_1)},$$

with $t_2 \geq t_1 > 0$ and

$$z_1 = (\log t_1 - \theta_0 - \theta_1^T \mathbf{w}) \exp(-\theta_2), \;\; z_2 = (\log t_2 - \theta_0 - \theta_1^T \mathbf{w}) \exp(-\theta_2).$$

As in Devroye (1986), this can be done by generating $U \sim U[0,1]$ and setting

$$\log L = \theta_0 + \theta_1^T \mathbf{w} + F_Z^{-1}\left(F_Z(z_1) + U\,[1 - F_Z(z_1)]\right) \exp\theta_2,$$

assuming that the inverse of $F_Z(z)$ exists and is either available in closed form or can be approximated to the required accuracy by the computer package of our choice.

3. $\{\tilde{C}_l \,|\, \mathbf{T}_l\}$: In the examples we shall consider we will obtain $\tilde{\mathbf{C}}_l$ by adding t to each element of $\mathbf{T}_l$, where t is the duration of stage II of the experiment. A more realistic censoring scheme would also take into account losses to follow–up prior to the end of the trial, with the censoring mechanism possibly dependent on the covariates.

4. $\{\tilde{\mathbf{C}}_m\}$: We shall assume that all elements of the future sample will be placed under observation at exactly the same time point, t_0 time units before the end of the trial. $\tilde{C}_m$ will then be degenerate at t_0, $0 \leq t_0 \leq t$. In clinical trials such an assumption is often unrealistic and an attempt should be made to model patient accrual probabilistically and to incorporate losses to follow–up during stage II.

5. $\{\theta \,|\, \mathbf{T}_d, \mathbf{T}_c, \mathbf{T}_l\}$: We will sample from the interim posterior of θ by running a Metropolis rejection independence chain, as described in Tierney (1994). If the envelope is in fact adequate, the method will produce independent draws from the exact posterior; otherwise the initial transient and the effect of dependence on standard errors will have to be taken into account.

5 Conducting a Sensitivity Analysis

Let us suppose that our client is a pharmaceutical company that will ultimately need to show to a regulatory agency that, in a trial in which at least s subjects were enrolled, administration of a new drug to the treatment group resulted in a median survival time at least r times as high as that of patients in the control group, with posterior probability in excess of a cutoff value p. To guard against possible side effects, which are of known severity for the standard treatment but not for the novel therapy, r will typically be strictly greater than unity with a whole range of values initially contemplated. Letting $g(\theta)$ denote the difference in the median loglifetimes between the two groups, the approximations in Papandonatos & Geisser (1997) require that the bulk of our computational efforts be directed towards calculating the first four derivatives at $g(\hat{\theta}_{n+m})$ of the terminal posterior of $g(\theta)$. Once these constants are in hand, the additional expense in letting r take multiple values will be minimal. This presents a clear advantage of our approach over that advocated by DiCiccio et al. (1990, 1991), when multiple tail areas need to be approximated. Similarly, we could vary p to see how sensitive our results are to requiring a more stringent level of confidence in our recommendations, at little further cost. A contour plot of the predictive probability of success as a function of p and r would then provide a nice visual supplement to our analysis. The above analysis assumes that both t, the duration of stage II and m, the size of the future sample, are fixed in advance with only p, r allowed to vary. But, there is no reason not to allow t, m to depend on the information gathered during stage I as well. Of most interest here would be the trade–off between increasing the length of the trial and

increasing the number of participants. For regression problems where the covariate settings can be chosen at will we obviously have an intricate design problem in our hands, which will not be dealt with here. When we cannot exercise any control over the covariates of the incoming participants, but can describe their probability distribution instead, we can try to design a factorial experiment with several (t, m) combinations, to which a response surface methodology could then be applied.

When m is fixed, a gradual increase in the time at which the imputed true lifetimes are being censored will lead to posterior modes that are not too far apart. These can then be used in succession as initial values for Newton's method. The same strategy can be used when we keep t fixed and slowly enlarge m. Since Newton's maximisation converges rapidly close to the maximum, but may encounter problems if a poor initial point is chosen, it is very encouraging to know that our starting values are likely to perform well.

6 An Example: Pike's Carcinogenesis Data

Table 1 about here

The data in table 1 come from a carcinogenicity experiment two groups of mice which has been reported in Pike (1966) and analysed extensively by Kalbfleisch & Prentice (1980). Since cancer occurs in a tissue when at least one of its cells becomes carcinogenic, Pike (1966) argued for a Weibull model for the time till cancer onset, due to the asymptotic derivation of the Weibull as the minimum of a large number of independent observations. The assumptions implicit in the analysis are that (i) death with tumor is equivalent to death from tumor, (ii) the tumor is rapidly lethal and (iii) eventually all mice will contract cancer, if they live long enough. As a result, mice that had not died at the time of the analysis or were tumor–free at death were regarded as censored observations. Following Kalbfleisch & Prentice (1980), we postulate a Weibull threshold model for the lifetimes in each group, and take the threshold and shape parameters to be common to the two groups, but allow the scale parameters to differ. If w were an indicator function denoting membership of the second group, our model would be equivalent to assuming that $[\lambda(L-\eta)/\rho^w]^\delta \sim \exp(1)$, where λ is the scale parameter of the first group and λ/ρ that of the second group. This would in turn imply that $\log(L-\eta)$ has a Gumbel$(\theta_0+\theta_1 w,\ \exp\theta_2)$ distribution where $\theta_0 = -\log\lambda$, $\theta_1 = \log\rho$ and $\theta_2 = -\log\delta$; a simple linear regression model to which the asymptotic approximations to marginal tail probabilities presented in Papandonatos & Geisser (1997) are applicable. Although their results can accommodate informative priors with ease, we will adopt a prior that is flat in the $\boldsymbol{\theta}$ parameterisation and which may be thought of as 'noninformative' in the sense that it reproduces conditional frequentist inferences under Type II censoring. It differs from Jeffreys' prior, which turns out to depend on the design matrix and which requires more detailed assumptions about the censoring mechanism.

The first group of mice has previously been analysed in a Bayesian predictive setting by Geisser (1993a), whose analysis was performed conditionally on $\eta = 100$ and $\delta = 3$. We will also condition on $\eta = 100$, but allow δ to be unrestricted, turning the problem into a bivariate one involving $\boldsymbol{\theta} = (\theta_0, \theta_2)$ alone. The additional information on θ_2 provided

by the second group will be ignored at this stage. This way our approximations for stage II censoring can be compared to his exact results for an uncensored experiment, at the additional expense of a one–dimensional numerical integration. The observation 188 appears twice in the first group, but its second occurrence was apparently omitted by Geisser. To make our results comparable to his, we will also exclude it from our initial analysis, keeping it in reserve for a perturbation analysis later on.

In order to illustrate the methodology outlined in earlier sections, we will first calculate the predictive probability that $L_{.50}$, the median lifetime of the first group of mice, will exceed 206 days with terminal posterior probability at least equal to p, if observations are taken on m additional mice for t days. The hypothesis $H_0 : L_{.50} \geq 206$ has an interim posterior probability of .907 and thus would be narrowly rejected in favour of $H_1 : L_{.50} < 206$ by any Bayesian hypothesis testing procedure requiring odds at least 10:1 in favour of H_0 for its acceptance.

Rather than focus on posterior odds alone, we may also want to impose the additional requirement that m and t should be chosen so as to virtually guarantee that a 99% equal tail terminal posterior interval for $L_{.50}$ will not exceed 45 days in length, down from 58.09 at the interim stage. The equal tail interval is chosen in preference to an H.P.D. region both for computational convenience and also in order to penalise extreme skewness in the posterior of $L_{.50}$, of the kind that could result from sampling plans that lead to very heavy censoring.

If we were to assume that the two mice censored at the interim stage were lost to follow up and that t was chosen long enough for all m future observations to die, the expressions in Geisser (1993a) could be manipulated to yield via univariate numerical integration the exact predictive probability $P[\, P[\, L_{.50} \geq 206 \,|\, \tilde{\mathbf{T}}_{n+m}] \geq p \,|\, \mathbf{T}_n]$ for various finite values of m and of the cutoff point p. The results are shown in fig. 1, as m is increased from 1 to 10^4 in powers of 10. We see that even a single additional observation introduces a lot of uncertainty, moving us from a point mass distribution at .907 to one that spreads its mass over the interval (0.2, 1). As the size of the future sample is increased further, the curves gradually get closer to the horizontal asymptote at .907, but convergence is rather slow.

Fig. 1 about here

A heuristic explanation for the limiting form of the predictive probability of eventual acceptance of H_0 can be given by first noting that, if the future sample size m were to increase without bounds, the distribution of the median lifetime $L_{.50}$ would tend to a point mass at some as yet unknown value. The best that can be said about this value at the interim stage is that it will exceed 206 days with probability .907. This would in turn suggest a discrete limiting predictive distribution for $P[\, L_{.50} \geq 206 \,|\, \tilde{\mathbf{T}}_{n+m}]$ supported entirely on $\{0, 1\}$ with weights .093 and .907 respectively. But then $P[\, P[\, L_{.50} \geq 206 \,|\, \tilde{\mathbf{T}}_{n+m}] \geq p \,|\, \mathbf{T}_n]$ should tend to .907 for all p in the interval (0, 1) in accord with result (2.5).

Fig 2 about here

In fig. 2 the value of p is increased from .65 to .95 in increments of .05, bringing out a fact obscured in fig. 1: the convergence to the asymptote need not be monotonic. Indeed, for low values of p it is not monotonic and, rather than gaining anything by taking a future sample of moderate size, we will probably end up worse off than if we took just a few

extra observations. On the other hand, for high values of p, significant gains are within our reach even for small values of m. To better understand why this happens, we should remind ourselves that the interim posterior probability of H_0 is .907 and any $p < .907$ would have resulted in acceptance of H_0 at the interim stage. It is then only natural that, by introducing uncertainty in the form of extra observations, we risk altering our conclusion in favour of H_1. Similarly, any $p > .907$ would have led to rejection of H_0 at the interim stage and continuing the experiment gives us a chance to obtain a more favourable result.

Although derived for an experiment with no second stage censoring, the exact results above are of some use to us in the case of t finite, since a decrease in the duration of the experiment can be thought of as equivalent to a reduction in the effective sample size. One can then identify a point in a particular curve, chose a fixed value of m and trace the predictive probability as m is decreased towards 0 to get a feel for the implications of a larger degree of censoring. For example, it is obvious from fig. 2 that an increase in the degree of censoring will have a much greater effect on the predictive probabilities at higher values of p, where there is more downswing potential.

If we assume that our resources allow a maximum of $m = 30$ mice to be added to the first treatment group, our aim could be to find the shortest censoring time consistent the requirement that a 99% terminal posterior interval for $L_{.50}$ be shorter than 45 days. This, in turn requires us to simulate future sample paths by drawing from the predictive density of $\tilde{\mathbf{T}}_{n+m} \mid \mathbf{T}_n$, which we will do by first getting a grip on the posterior of $\boldsymbol{\theta}$.

Since $p(\theta_0 \mid \theta_2, \mathbf{T}_n)$ is given in Geisser (1993a), in order to construct an adequate envelope to $p(\theta_0, \theta_2 \mid \mathbf{T}_n)$ we only need to a good approximation to $p(\theta_2 \mid \mathbf{T}_n)$. Papandonatos & Geisser (1997) have shown that taking $(\theta_2 + 1.227)/.548$ to be distributed as the logarithm of a $F(827.524, 16.053)$ variable results in approximation indistinguishable from the true density to within plotting accuracy. Five hundred simulated $\boldsymbol{\theta}$ values are given in fig. 3, with .8, 0.5, .2, .01 and .001 contours of their exact scaled posterior superimposed. If we assume a $\chi^2_{(2)}$ approximation to -2 log $\left[p(\boldsymbol{\theta} \mid \mathbf{T}_n)/p(\hat{\boldsymbol{\theta}} \mid \mathbf{T}_n)\right]$, the γ'th contour of the scaled posterior corresponds to an approximate $1-\gamma$ joint credible interval. We thus conclude that our sample is evenly spread over the range of the parameter space receiving non-negligible posterior mass.

Fig. 3 about here

In order to generate samples from the joint predictive density of the lifetimes of the future sample, we first notice that - conditional on $\boldsymbol{\theta}$ - the individual lifetimes are independent Weibull-distributed random variables. Therefore, they are exchangeable unconditionally and have the same marginal distribution, which can be approximated by a finite mixture of Weibull distributions evaluated at the simulated parameter pairs. The approximation is given in table 2 and reveals that there is no real need to consider experiments longer than twelve months in duration, since almost all the mice will have died by then. On the other hand, it seems that four months constitute the minimum duration required for obtaining additional information about the death times from the future sample. The expected number of deaths at any given censoring time does not, however, convey the whole picture, with censored observations possibly quite informative for the lower quantiles of the lifetime

Table 2 about here

distribution, but less so for the median. So, we also need to examine the simulated terminal posteriors of $L_{.50}$ directly and estimate selected quantiles of the predictive distribution of the length of a 99% terminal credible interval.

This brings us to the most time–consuming part of the simulation: the maximisation of the five hundred simulated terminal posteriors, each one based on thirty additional lifetimes which are first truncated at t and then combined with the initial sample. The maxima for four selected second stage durations t are plotted in fig. 4. A comparison of figs. 3 and 4 shows a tendency for the maxima to shrink the simulated parameter pairs used in generating the m future lifetimes towards the mode $\hat{\theta}_n$ of the interim posterior, the degree of shrinkage depending on the length of phase II. The longer t is, the more we learn about the simulated value $\tilde{\theta}_m$ and the closer the terminal posterior mode moves towards it. Even for a fixed censoring time, it is apparent from the plots that the degree of shrinkage varies between the axes of the ellipse defined by the normal approximation to the interim posterior of θ. An intuitive explanation is not hard to find: if we were to work in terms of centered and asymptotically uncorrelated coordinates, we would find them to be proportional to proportional to the .7605 and .0001 quantiles of the loglifetimes. What the plots then tell us is that, as the degree of censoring increases, the terminal posterior mode of $L_{.7605}$ is much closer to the interim mode than that of $L_{.0001}$, which is pulled towards its simulated value. This makes sense, since heavy censoring allows to us to glean little new information about the seventy sixth percentile of the lifetime distribution, but quite a lot about the minimum possible death time.

Fig. 4 about here

This observation argues against the use of convex combinations of $\hat{\theta}_n$ and $\tilde{\theta}_m$ in initialising the Newton iteration. We found that a a matrix weighted average of $\hat{\theta}_n$ and $\tilde{\theta}_m$ with weights given by the information in the likelihoods of the initial and future samples respectively performed better in this regard. When multiple censoring times are contemplated, all quite close to each other, a possible initialisation strategy would be to start the maximisations of each simulated terminal posterior at the smallest censoring time with the guess suggested above and then use the new mode as the best guess at the mode of the posterior that results when truncating the same imputed lifetimes at the immediately higher censoring time. In general we found that Newton's method converged in 3–5 iterations at the lowest censoring time and 1–2 iterations as t was subsequently increased by one month at a time.

e

Once the posterior modes are in hand, the .005 and .995 quantiles of the normal and Laplace approximations to the marginal of $L_{.50}$ can be calculated with little additional computational effort. Their difference can be used in turn to calculate the predictive distribution of 99% terminal credible interval lengths, whose quartiles are shown in fig. 5. We immediately notice that the normal approximation has interquartile range similar to that of the Laplace approximation, but its median is lower, the difference increasing with the degree of censoring. The reason that the length distribution converges to a point mass at around 120 days is that, with the expected number of uncensored observations close to 0,

Fig. 5 about here

all simulated terminal posteriors are almost identical.

As for the probability of obtaining a terminal 99% interval no longer than 45 days, we present it in table 3. The results are very interesting, because they clearly show how misleading the normal approximation to the posterior of $L_{.50}$ can be for a sample size even as large as 48, when the degree of censoring is high. In particular, the normal approximation results would encourage us to curtail the experiment at six months, whereas the Laplace approximation indicates that a minimum of nine months is required for us to have a high degree of confidence in achieving the required interval length. It could be argued that early stopping is not of too strong a concern to a Bayesian, who, unlike a frequentist, may take multiple looks at the data without having to adjust the length of the credible intervals. Suppose, however, that the problem was not one of choosing between different values of t for the same m, but of varying m for a fixed t; then the inadequacies of the normal approximation would be more readily apparent: by misleading us into thinking that observing 30 mice over six months would be adequate, we would end up with an insufficiently informative experiment. If we then attempted to save the experiment by taking yet another sample of mice, we would have to wait at least another four months to observe any deaths and possibly seriously overrun our time schedule.

Table 3 about here

In order to conduct a sensitivity analysis similar to the one in Geisser (1993a), we varied p from 0.65 to 0.95 in steps of 0.05 and obtained tables 4, 5, which approximate the predictive probability that $P[\, L_{.50} \geq 206 \,|\, \tilde{\mathbf{T}}_{48}] \geq p$ with a standard error not exceeding 0.02. Their last column pertains to an uncensored experiment and was calculated using the results in Geisser (1993a).

Tables 4, 5 about here

A cursory comparison of the entries in tables 4, 5 shows that use of the normal approximation results in mild overoptimism in ascertaining the probability of success at censoring times higher than seven months, while the picture is reversed for censoring times of seven months or lower. The discrepancy is not as remarkable as the one previously obtained for the distribution of interval lengths; it increases with p but never exceeds 6%. [3] It is also seen that, that for $p \leq .95$, the odds of achieving our goal are at least 2:1 in our favour for any experiment longer than seven months in duration and that an increase in the length of the experiment from seven to twelve months brings only a negligible improvement in the odds of success. So, the critical factor in fixing the length of the second phase should be the probabi lity of obtaining a sufficiently informative sample, as determined by table 3.

It is interesting that, even though the interim probability of H_0 is just above .90, the predictive probability that this will be the case at the end of the more informative experiment is only .79. Still, there are sufficient grounds for being optimistic that we will be able to decide in favour of H_0 at the end of the trial, for all values of p under consideration.

[3] When using the tables to estimate changes in the predictive probabilities as either p or t vary, it should be kept in mind that entries in the table should be positively correlated not only within the same row but also within columns, because the same future lifetimes were truncated at all eight censoring points to yield the simulated times under observation. This can be seen as an elementary attempt at variance reduction using common variates.

Finally, we examine the changes in the predictive probabilities if the extra observation at 188 days is included in the original sample and the two mice censored at the interim stage are assumed to have continued under observation, instead of being lost to follow–up. The predictive densities of the true lifetimes of the censored observations can be derived as mixtures of left–truncated Weibull distributions and suggest that the probability that either one observation is still surviving at the end of stage II is minute. Rather than repeating the time–consuming maximisations of the new terminal posteriors, one can find the new predictive probabilities by perturbing the original simulated posteriors, as outlined in Papandonatos & Geisser (1997). In addition to including the extra lifetime, this involves dropping the two censored observations, imputing their lifetimes and then recensoring them according to the length of stage II. The results are given in tables 6, 7. Apart from the row corresponding to $p = .95$, the entries of table 6 vary little from those of table 5. More significant is the increase in the informativeness of the experiment, as manifested by the higher probabilities of achieving a 99% terminal credible interval length of 45 days, whatever the censoring time. If our rule of the thumb is to set the censoring time to the smallest value consistent with a predictive probability of obtaining the desired interval length of at least .9, table 7 suggests that we can reduce t from 240 to 225 days if the extra three mice became available to observation during the final stage of the experiment, while maintaining the probability that $P[\, L_{.50} \geq 206 \,|\, \tilde{\mathbf{T}}_{19}] \geq p$ to within 2% of its original value at 240 days.

Tables 6, 7 about here

Although the single group example is useful in that it allows us to compare the simulation results for stage II censoring to exact ones obtained for an uncensored experiment, a typical application of our methodology would involve the comparison of two treatments in terms of a parameter indicating treatment effectiveness, a setup not previously examined within a fully Bayesian predictive setting.

In the Gumbel regression model we introduced earlier in this section, $\rho = \exp(\theta_1)$ is the ratio of the median lifetime in excess of 100 days for the second group to that of the first group and can thus be used as a criterion for comparing the two treatments. At the time of the interim analysis, this criterion would strongly favour the second treatment. Before pronouncing in its favour, however, we should note that two treatments are often considered equivalent if the ratio of the median lifetimes of the two groups falls with high posterior probability within a short interval centered at one. This approach is common in bioequivalence studies, in which a new drug formulation is being compared with an established standard with well–understood side–effects; one is understandably reluctant to abandon the better–known drug unless its competitor is demonstrably superior. The interim odds in favour of the second treatment regime when the equivalence zone is of the form $(1-h,\, 1+h)$ are given in table 8 for some common choices of h. They are calculated by dividing the probability of exceeding the upper boundary of the equivalence zone by the probability of falling below it, using interim data on both groups of mice. Larger values of h correspond to increased conservatism on the part of the investigator, who may prefer to withhold judgment in either direction until additional information becomes available on the

Table 8 about here

two treatments. As can be found from the table, the interim probability that ρ exceeds one is .962, but that of ρ exceeding 1.20 is only .601, so that the introduction of an equivalence zone of 20% makes the choice between the two treatments much less clearcut.

We will now assume that we require odds at least 10:1 in favour of the second treatment, with the treatments declared equivalent if the ratio of the median lifetimes in excess of 100 days are within 10% of each other. Since no decision can then reached at the interim stage, it is of interest to us to predict whether observing an additional 60 mice for periods up to one year is likely to allow us to reach a conclusive result. Assuming that our resources do not allow larger scale experimentation, the trial would have to be abandoned if that predictive probability was small. Equalizing the sample sizes of the two groups produces a predictive distribution for the terminal odds in favour of the second treatment whose quantiles are given in table 9. Among the second stage durations we examined, the chances of obtaining odds of at least 10:1 in favour of the second treatment are better than even for all durations greater than seven months. However, although the predictive distribution rapidly acquires a long right tail, its lower quantiles also move away from the interim value of 27:5, with the first quartile eventually stabilising at odds of around 7:2 after six months. This seems to imply that although there is considerable upswing potential, one must also accept the possibility that the final evidence in favour of the null hypothesis will be weaker than at the interim stage. In this sense, a prolongation of the trial does not allow us to sample to a foregone conclusion, but simply reflects an informed decision on our part to commit additional resources to demonstrating the effectiveness of a treatment regime we regard as sufficiently promising.

Table 9 about here

As can be seen from the credible intervals for the log–odds ratio traced out in fig. 6, both the variance and the right skewness of the predictive distribution increase with time. A strategy aimed at maximising the probability of accepting the hypothesis that the second is superior would then be to prolong the second stage duration for as long as it takes for all the additional mice to die, i.e. for approximately one year. Indeed, we know from result (2.5) that, for a large enough experiment, the predictive probability of obtaining any positive odds in favour of the second treatment should tend to $P[\rho \geq 1.10 \,|\, \boldsymbol{T}_{40}] = .843$. As can be seen from table 10, an additional sample of 60 mice is not large enough to get us close to this limiting value, but does approach it from below as the effective sample size increases, until it stabilises at around .60 for the essentially uncensored experiment of a one–year duration. Getting any closer to the limiting probability than that requires increasing the number of mice allocated to the two treatments, rather than simply prolonging the second stage of the experiment. Still, to the extent that the intuition gained from fig. 2 is relevant to the current setting, it is likely that most of the gains from taking an additional sample have already been exhausted and that raising the predictive probability to .85 would require prohibitive increases in the size of the future sample.

Fig. 6 about here

Table 10 about here

7 Conclusions / Future Work

In this paper we showed how a combination of simulation and analytic approximation can aid Bayesian interim analyses of Weibull–distributed lifetime data. Although our examples dealt only with right censoring, extensions to left and interval censoring are trivial to implement. The required expansions are also easy to obtain for other error distributions, but their adequacy needs to be examined case–by–case. The Laplace approximation may suffer when the likelihood is not strictly unimodal or the prior is in serious conflict with the data.

The bulk of the computation effort required for our analysis is directed towards the maximisation of the simulated terminal posteriors. Since the convergence of Newton's method is quadratic in a neighbourhood of the mode and occurs in a single step when the contours are elliptical, good initial estimates of the mode and parameterisations that lead to near–normal posteriors can drastically reduce computation time. More experience with problems of this kind may allow us to suggest starting values better than the matrix weighted average that we employed. Even more promising are variance reduction techniques that allow us to obtain precise estimates of the predictive probabilities of reaching our goal using only a small number of simulated terminal posteriors.

Alternative parameterisations should also be considered, with their effect on the simulations, maximisations and analytic expansions assessed separately. There is no reason why the same parameterisation should be optimal for all of the above, though globally log–concave likelihoods with nearly elliptical contours are always a boon.

Other design criteria may also be entertained. When the variance of the parameter denoting treatment effectiveness increases with its mean, relative rather than absolute length of terminal posterior intervals could be used in designing the experiment. Alternatively, our analysis could focus on the posterior probability of the hypothesis of interest alone, calculated under a range of priors ranging from the optimistic to the sceptical, as recommended by Spiegelhalter & Freedman (1994). This assumes that we neither engage in continuous monitoring of the data nor have freedom in setting the levels of the covariates. Too much latitude in this respect would result in a huge increase of the computational burden, that cannot be realistically handled at this point in time, but may be more readily entertained in the near future, with the advent of massively parallel computers.

In summary, the promise held by Bayesian interim analysis of regression models with possible censoring has not yet been fully realised, but the results to date seem sufficiently encouraging to justify further research in this area.

8 References

Armitage, P. (1989). "Discussion of the paper by Jennison & Turnbull", *Journal of the Royal Statistical Society, Series B*, **51**, 333–335.

Berry, D.A. (1985). "Interim analyses in clinical trials: Classical vs. Bayesian Approaches", *Statistics in Medicine*, 4, 521–526.

Bernardo, J.M. (1979). "Expected information as expected utility", *The Annals of Statistics*, **7**, 686–690.

Choi, S.C., and Pepple, P.A. (1985). "Monitoring clinical trials based on predictive probability of significance", *Biometrics*, **45**, 317–323.

DeMets, D.L., and Lan, K.K.G. (1984). "An overview of sequential methods and their application in clinical trials", *Communications in Statistics, Part A–Theory and Methods*, **13**, 2315–2338.

Devroye, L. (1986). *Non–uniform random variate generation.* Springer–Verlag, New York.

DiCiccio, T.J., and Field, C.A. (1991). "An accurate method for approximate conditional and Bayesian inference about linear regression models from censored data", *Biometrika*, **78**, 903–910.

DiCiccio, T.J., Field, C.A., and Fraser, D.A.S. (1990). "Approximations of marginal tail probabilities and inference for scalar parameters", *Biometrika*, **77**, 77–95.

Doob, J. L. (1949). "Le Calcul des Probabilités et ses Applications.", *Colloques Internationaux du Centre National de la Recherche Scientifique*, 23–27.

Doob, J. L. (1953). *Stochastic Processes.* John Wiley & Sons, New York.

Geisser, S. (1992). "On the curtailment of sampling", *The Canadian Journal of Statistics*, **20**, 297–309.

Geisser, S. (1993a). "Bayesian interim analysis of censored exponential observations", *Statistics & Probability Letters*, **18**, 163-168.

Geisser, S. (1993b). *Predictive Inference: An Introduction.* Chapman & Hall, London.

Geisser, S., and Johnson, W. (1993). "Interim analysis for normally distributed observables", *Multivariate Analysis and Its Applications*, IMS Lecture Notes – Monograph Series.

Kalbfleisch, J.D., and Prentice, R.L. (1980). *The statistical analysis of failure time data.* John Wiley & Sons, New York.

Kass, R.E., Tierney, L., and Kadane, J.B. (1990). The validity of posterior expansions based on Laplace's method. *Bayesian and likelihood methods in Statistics and Econometrics* (eds S. Geisser, J.S. Hodges, S.J. Press, and A. Zellner), pp. 472–488. Elsevier Science Publishers, Amsterdam.

Lindley, D.V. (1956). "On a measure of the information provided by an experiment", *Annals of Mathematical Statistics*, **27**, 986–1005.

Papandonatos, G.D., and Geisser, S. (1997). "Laplace Approximations for Censored Linear Regression Models", *The Canadian Journal of Statistics*, to appear.

Parmigiani, G., and Berry, D.A. (1994). "Applications of Lindley Information Measure to the design of clinical experiments", In P.R. Freeman and A.F.M. Smith, eds., *Aspects of Uncertainty* , 329–348. John Wiley & Sons, New York.

Pike, M.C. (1966). A method of analysis of a certain class of experiments in carcinogenesis.

Spiegelhalter, D.J., and Freedman, L.S. (1988). "Bayesian approaches to Clinical Trials". In J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, eds., *Bayesian Statistics*, **3**, 453–477. Clarendon Press, Oxford.

Spiegelhalter, D.J., Freedman, L.S., and Blackburn, P.J. (1986). "Monitoring clinical trials: conditional or predictive power?", *Controlled Clinical Trials*, **7**, 8–17.

Spiegelhalter, D.J., Freedman, L.S., and Parmar, M.K.B. (1994). "Bayesian approaches to randomised trials", *Journal of the Royal Statistical Society, Series A*, **157**, 357–416.

Tierney, L. (1994). "Markov Chains for exploring Posterior Distributions" (with discussion), *The Annals of Statistics*, **22**, 1701–1762.

Tierney, L., and Kadane, J.B. (1986). "Accurate approximations for posterior moments and marginal densities", *Journal of the American Statistical Association*, **81**, 82–86.

Viveros, R., and Sprott, D.A. (1987). "Allowance for skewness in maximum likelihood estimation with application to the location-scale model", *Canadian Journal of Statistics*, **15**, 349–361.

Table 1: Days to vaginal cancer mortality in mice after carcinogenic insult.

Group 1	143,	164,	188,	188,	190,	192,	206,	209,	213,	216,	
	220,	227,	230,	234,	246,	265,	304,	216*,	244*		
Group 2	142,	156,	163,	198,	205,	232,	232,	233,	233,	233,	233,
	239,	240,	261,	280,	280,	296,	296,	323,	204*,	344*	

Data from Pike (1966). Asterisks denote censored observations.

Table 2: Predictive distribution of the lifetime L of a future observation.

l	120	150	180	210	225	240	270	360
$P[L \leq l \mid \mathbf{T}_{18}]$	.004	.049	.179	.409	.544	.674	.869	.996

Table 3: Probability of meeting design criterion when mice are lost to follow-up.

Approximation	Length of Phase II in Days							
	120	150	180	210	225	240	270	360
Normal	.00	.00	.95	.98	.97	.97	.98	.96
Laplace	.00	.00	.37	.82	.88	.92	.94	.95

Table 4: $P[\, P[\, L_{.50} \geq 206 \mid \tilde{\mathbf{T}}_{48}] \geq p \mid \mathbf{T}_{18}]$: Normal Approximation.

p	Length of Phase II in Days								
	120	150	180	210	225	240	270	360	∞
.65	.98	.96	.94	.93	.92	.92	.92	.92	.90
.70	.98	.94	.92	.91	.90	.91	.91	.91	.89
.75	.97	.92	.89	.89	.89	.89	.89	.89	.88
.80	.89	.88	.87	.87	.87	.86	.86	.86	.85
.85	.89	.83	.83	.83	.83	.84	.84	.84	.82
.90	.89	.71	.74	.78	.79	.80	.81	.81	.78
.95	.00	.57	.61	.67	.70	.73	.74	.74	.72

Table 5: $P[\, P[\, L_{.50} \geq 206 \mid \tilde{\mathbf{T}}_{48}] \geq p \mid \mathbf{T}_{18}]$: Laplace Approximation.

p	Length of Phase II in Days								
	120	150	180	210	225	240	270	360	∞
.65	.98	.97	.95	.93	.92	.92	.92	.92	.90
.70	.98	.95	.93	.91	.90	.91	.91	.91	.89
.75	.97	.92	.90	.89	.89	.89	.89	.89	.88
.80	.89	.91	.88	.87	.86	.86	.86	.86	.85
.85	.89	.84	.84	.84	.83	.83	.82	.82	.82
.90	.89	.74	.76	.79	.79	.79	.79	.79	.78
.95	.00	.58	.63	.67	.68	.68	.68	.68	.72

Table 6: $P[\,P[\,L_{.50} \geq 206 \,|\, \tilde{\mathbf{T}}_{49}] \geq p \,|\, \mathbf{T}_{19}]$: Laplace Approximation.

p	Length of Phase II in Days							
	120	150	180	210	225	240	270	360
.65	.98	.94	.93	.92	.91	.91	.91	.91
.70	.95	.93	.91	.90	.89	.89	.90	.90
.75	.93	.91	.89	.87	.87	.87	.87	.87
.80	.90	.87	.85	.85	.85	.85	.85	.84
.85	.88	.81	.81	.82	.82	.82	.82	.82
.90	.51	.70	.75	.76	.77	.78	.77	.78
.95	.00	.48	.63	.66	.67	.65	.64	.64

Table 7: Probability of meeting design criterion when no mice are lost to follow–up.

Length of Phase II in Days	120	150	180	210	225	240	270	360
Laplace Approximation	.00	.22	.67	.86	.92	.93	.96	.97

Table 8: Interim odds in favour of treatment 2 vs. half–length of equivalence zone.

h	.20	.15	.10	.05	.01	.00
$\frac{P[\,\rho \geq 1+h \,\vert\, \mathbf{T}_{40}]}{P[\,\rho < 1+h \,\vert\, \mathbf{T}_{40}]}$	1.51	2.78	5.38	11.18	21.23	25.12

Table 9: Predictive distribution of the terminal odds in favour of treatment 2.

Quantile	Length of Phase II in Days								
	120	150	180	210	240	270	300	330	360
.10	2.79	2.44	1.86	1.62	1.08	1.07	1.09	1.01	0.99
.25	4.33	3.91	3.50	3.53	3.66	3.55	3.62	3.53	3.49
.50	6.36	6.72	7.14	9.71	13.21	16.54	17.84	20.72	21.92
.75	8.79	10.57	15.25	29.27	61.93	110.83	157.64	167.68	171.69
.90	11.93	16.56	32.84	119.30	419.30	1055.74	1718.76	1698.49	1944.89

Table 10: Predictive probability that treatment 2 will be judged superior at termination.

Length of Phase II in Days	120	150	180	210	240	270	300	330	360
$P\left[\frac{P[\,\rho \geq 1.10 \,\vert\, \tilde{\mathbf{T}}_{100}]}{P[\,\rho < 1.10 \,\vert\, \tilde{\mathbf{T}}_{100}]} \geq 10 \,\middle\vert\, \mathbf{T}_{40}\right]$	.18	.28	.38	.49	.55	.58	.59	.60	.60

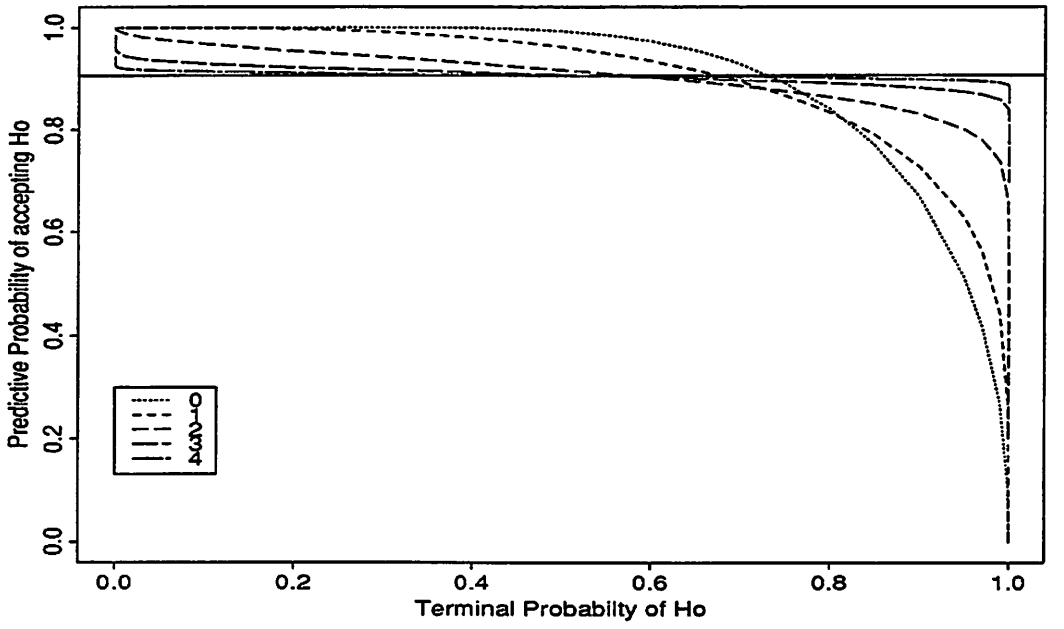

Figure 1: $P[\,P[\,L_{.50} \geq 206 \,|\, \tilde{\mathbf{T}}_{18+m}] \geq p \,|\, \mathbf{T}_{18}]$ vs. p for $\log_{10} m = 0, 1, 2, 3, 4$

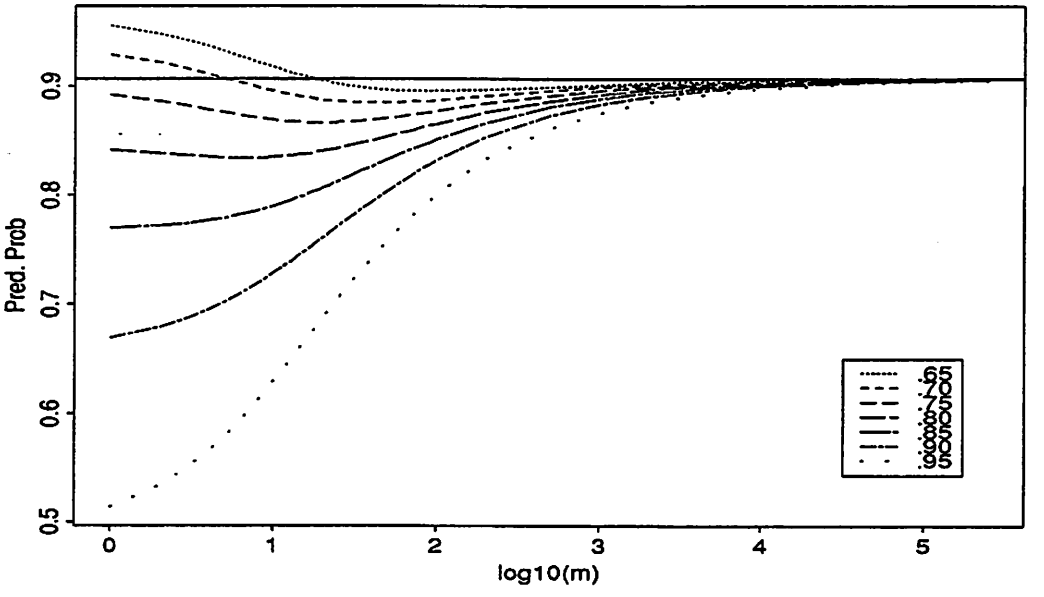

Figure 2: $P[\,P[\,L_{.50} \geq 206 \,|\, \tilde{\mathbf{T}}_{18+m}] \geq p \,|\, \mathbf{T}_{18}]$ vs. $\log_{10} m$ for $p = .65, \ldots, .95$ in steps of .05

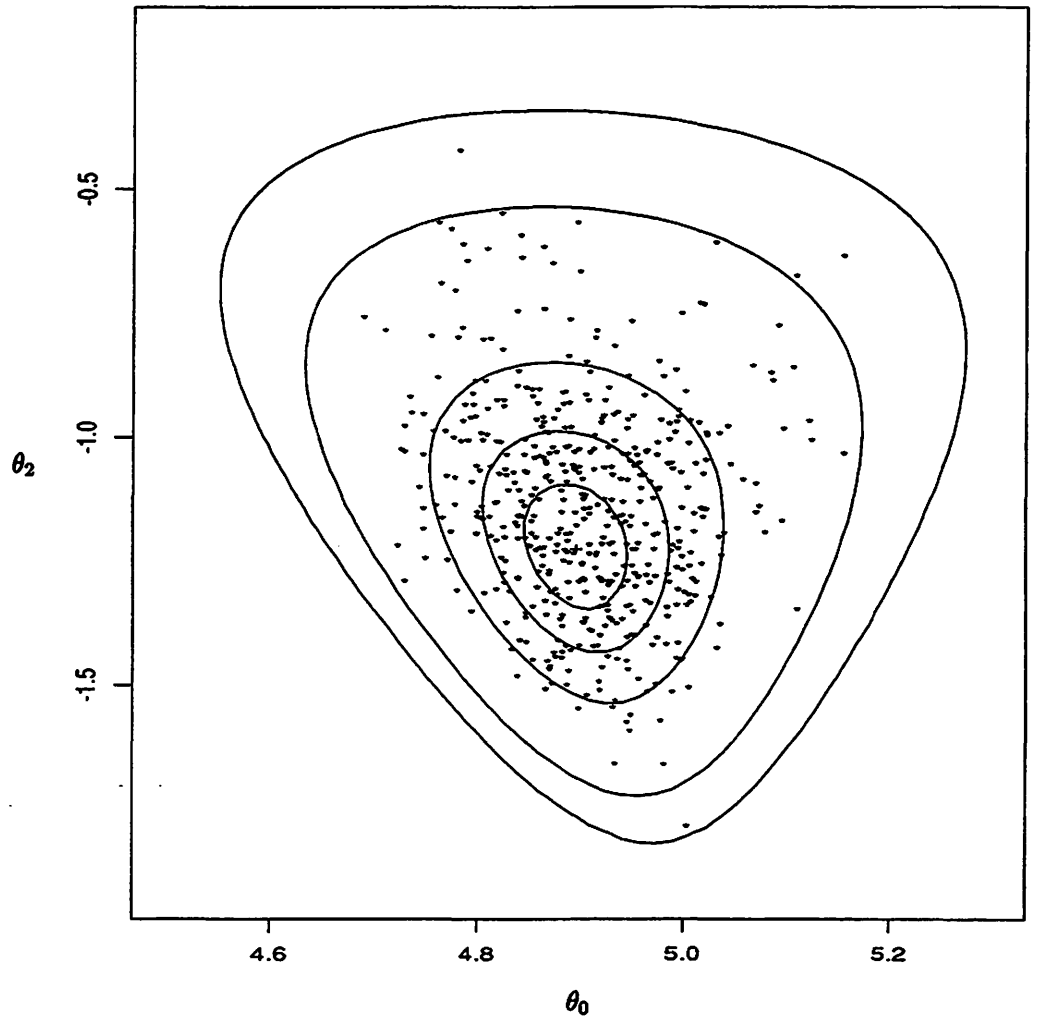

Figure 3: The .8, .5, .2, .01 and .001 contours of the interim posterior of θ with 500 posterior draws superimposed.

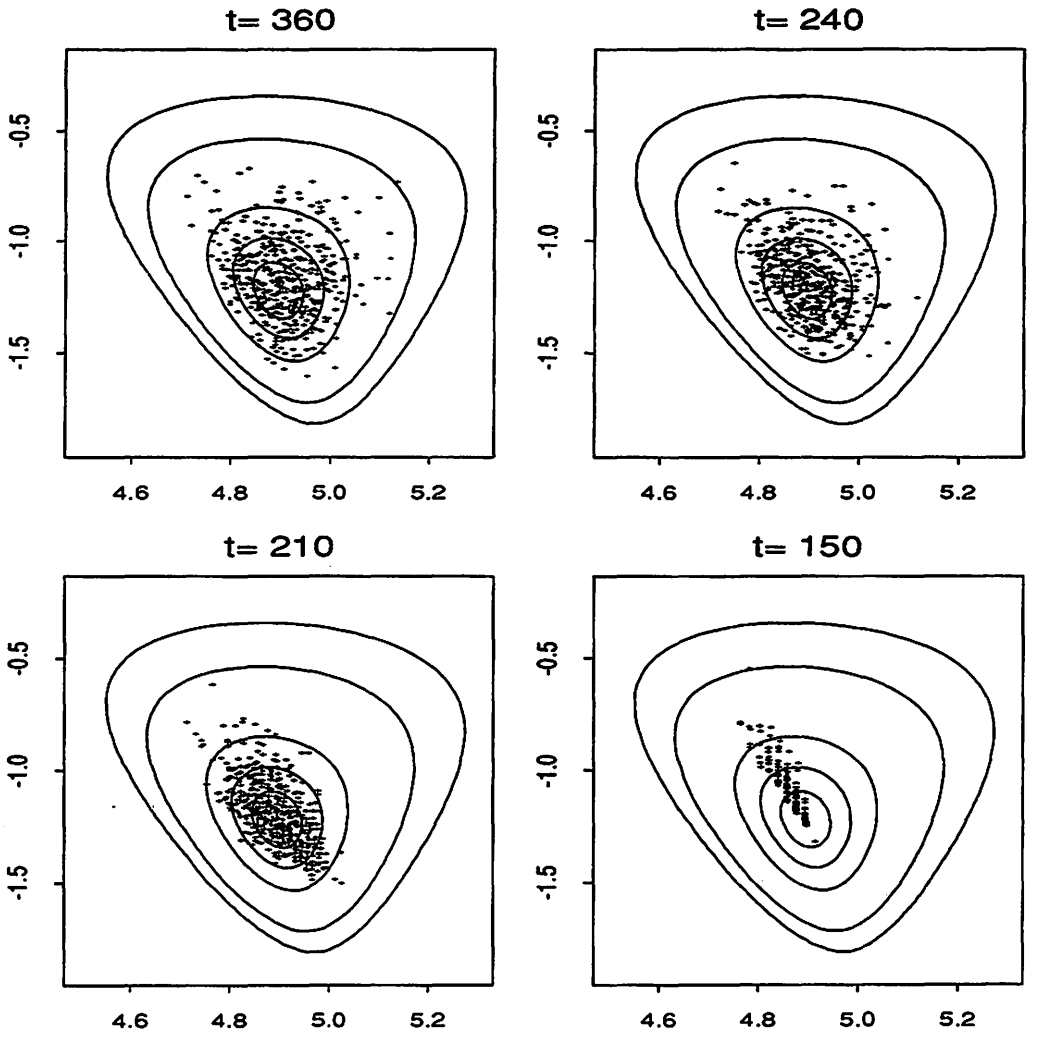

Figure 4: The .8, .5, .2, .01, and .001 contours of the interim posterior of θ with modes of 500 simulated terminal posteriors superimposed. Stage II durations given by t.

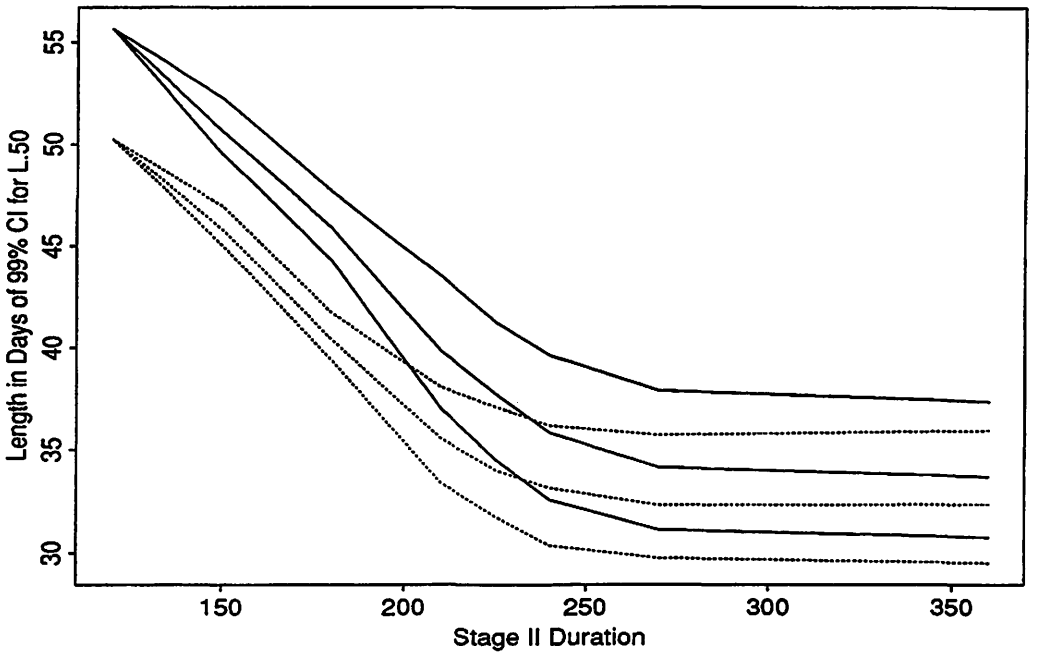

Figure 5: Quartiles of the predictive distribution of the length of a 99% terminal credible interval for $L_{.50}$ vs. stage II duration: Laplace Approx. —, Normal Approx. - - -.

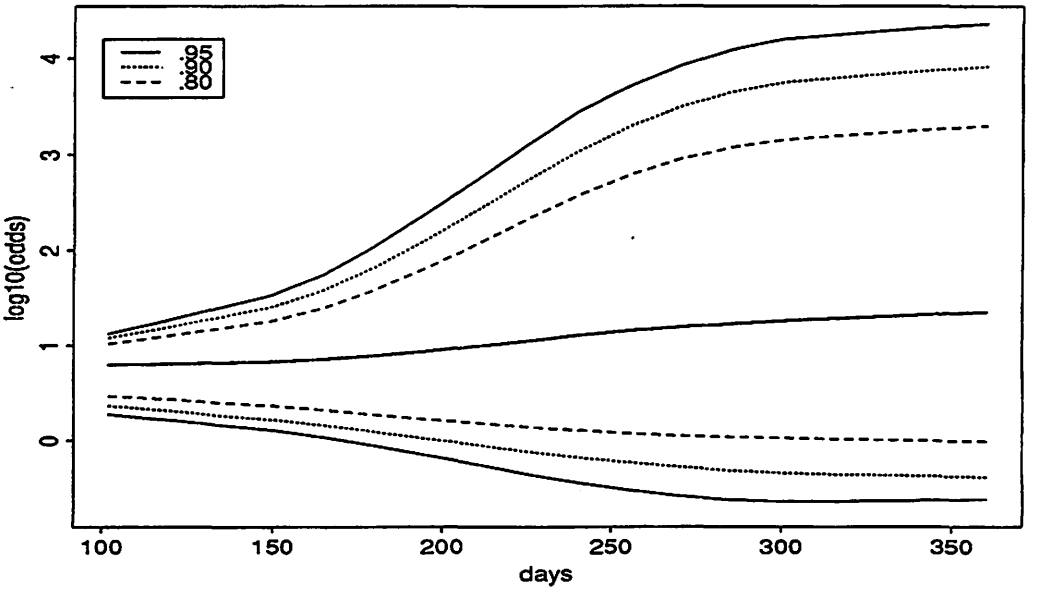

Figure 6: Credible intervals of the predictive distribution of $\log_{10} \frac{P[\rho \geq 1.1 \,|\, \tilde{\mathbf{T}}_{100}]}{P[\rho < 1.1 \,|\, \tilde{\mathbf{T}}_{100}]}$ vs. length of Phase II. Heavy line is the median.